

Misassembly Detection using Paired-End Sequence Reads and Optical Mapping Data

Martin D. Muggli¹, Simon J. Puglisi², Roy Ronen³, and Christina Boucher¹

¹ Department of Computer Science,
Colorado State University, Fort Collins, CO.
{muggli,cboucher}@cs.colostate.edu

² Department of Computer Science,
University of Helsinki, Finland
puglisi@cs.helsinki.fi

³ Department of Computer Science and Engineering,
University of California, San Diego
rronen@cs.ucsd.edu

Abstract. A crucial problem in genome assembly is the discovery and correction of misassembly errors in draft genomes. We develop a method that will enhance the quality of draft genomes by identifying and removing misassembly errors using paired short read sequence data and optical mapping data. We apply our method to various assemblies of the loblolly pine and *Francisella tularensis* genomes. Our results demonstrate that we detect more than 54% of extensively misassembled contigs and more than 60% of locally misassembled contigs in an assembly of *Francisella tularensis*, and between 31% and 100% of extensively misassembled contigs and between 57% and 73 % of locally misassembled contigs in the assemblies of loblolly pine. MISSEQUEL can be downloaded at <http://www.cs.colostate.edu/seq/>.

1 Introduction

Comparing genetic variation between and within a species is a fundamental activity in biological research. For example, there is currently a major effort to sequence entire genomes of agriculturally important plant species to identify parts of the genome variable in a given breeding program and, ultimately, create superior plant varieties. Robust genome assembly methods are imperative to these large sequencing initiatives and other scientific projects [1–4] because scientific analyses frequently use those genomes to determine genetic variation and associated biological traits.

At present, the majority of assembly programs are based on the Eulerian assembly paradigm [5, 6], where a de Bruijn graph is constructed with a vertex v for every $(k - 1)$ -mer present in a set of reads, and an edge (v, v') for every observed k -mer in the reads with $(k - 1)$ -mer prefix v and $(k - 1)$ -mer suffix v' . A contig corresponds to a non-branching path through this graph. We refer the reader to Compeau et al. [7] for a more thorough explanation of de Bruijn graphs and their use in assembly. The assemblers Euler-SR [8], Velvet [9], SOAPdenovo [10], ABySS [11] and ALLPATHS [12] all use this paradigm and follow the same general outline: extract k -mers from the reads, construct a de Bruijn graph from the set k -mers, simplify the graph, and construct contigs.

One crucial problem that persists in Eulerian assembly (and genome assembly, in general) is the discovery and correction of misassembly errors in draft genomes. We define a *misassembly error* as an assembled region that contains a significantly large insertion, deletion, inversion, or rearrangement that is the result of decisions made by the assembly program. Identification of misassembly errors is important because true biological variations manifest in similar ways and thus, these errors can be easily misconstrued as true genetic variation [13]. This can mislead a range of genomic analyses. We

note that the exact definition of a misassembly error can vary, and adopt the standard definition used by Quast [14] and other tools. See section 3.1 for this exact definition. Once the existence and location of a misassembly is identified, it can be removed by segmenting the contig at that location.

We present a computational method for identifying misassembly errors using a combination of short reads and optical mapping data. Optical mapping is a system developed in 1993 [15] that can construct ordered, genome-wide, high-resolution restriction maps. The system works as follows [16, 17]: an ensemble of DNA molecules adhered to a charged glass plate are elongated by fluid flow. An enzyme is then used to cleave them into fragments at loci where the corresponding recognition sequence occurs. Next, the fragments are highlighted with fluorescent dye and imaged under a microscope. Finally, these images are analyzed to estimate the fragment sizes, producing a molecular map. Since the fragments stay relatively stationary during the aforementioned process, the images capture their relative order and size [18]. Multiple copies of the genome undergo this process, and a consensus map is formed that consists of an ordered sequence of fragment sizes, each indicating the approximate number of bases between occurrences of the recognition sequence in the genome [19].

Although optical mapping data has been used for discerning structural variation in the human genome [20], and for scaffolding and validating contigs for several large sequencing projects — including those for various prokaryote species [21–23], rice [24], maize [25], mouse [26], goat [27], parrot [28], and *Amborella trichopoda* [29] — there exists no publicly available tools for using this data for misassembly correction using short read and optical mapping data.

Our tool, which we call MISSEQUEL, predicts which contigs are misassembled and the approximate locations of the errors in the contigs. It takes as input the paired-end sequence read data, contigs, an ensemble of optical maps, and the restriction enzymes used to construct the optical maps. MISSEQUEL first uses the paired-end read data to divide the contigs into two sets: those that are predicted to be correctly assembled and those that are not. Then the set of contigs that are candidates for containing misassembly errors are further divided into misassembled contigs and correctly assembled contigs using optical mapping data. Fundamental to the first step is the concept of a *red-black positional de Bruijn graph*, which encapsulates recurring artifacts in the alignment of the sequence read data to the contigs and their position in the contig. The red vertices in this graph indicate if a contig is likely to be misassembled and also flag the location where the misassembly error occurs. These locations are called *misassembly breakpoints*.

In the second stage of MISSEQUEL where optical mapping data is used, the contigs conjectured to be misassembled are *in silico* digested with the set of input restriction enzymes and aligned to the optical map using Twin [30]. Based on the presence or absence of alignment, a prediction of misassembly is made. The *in silico* digestion process computationally mimics how each restriction enzyme would cleave the segment of DNA defined by the contig, returning “mini-optical maps” that can be aligned to the optical map for the whole genome. An important aspect of our work is that it highlights the need to use another source of information, which is independent of the sequence data but representative of the same genome, in order to identify misassembly errors. We show that optical mapping data can be used as this information source.

We give results for the *Francisella tularensis* and loblolly pine genomes. Each genome was assembled using various de Bruijn graph assemblers and then misassembly errors were predicted. Our results on *Francisella tularensis* show that MISSEQUEL correctly identifies (on average) 86% and 80% of locally and extensively misassembled contigs, respectively. This is a considerable improvement on existing methods, which identified (on average) 26% and 16% of locally and extensively

misassembled contigs, respectively, in the same assemblies. The results on the loblolly pine genome assemblies show similar improvement. Lastly, our results demonstrate we are capable of significantly decreasing the false positive rate in all assemblies of *Francisella tularensis* and loblolly pine by incorporating optical map data into the prediction; the reduction was between 29% and 74%.

Related Work. Both amosvalidate [31] and REAPR [32] are capable of identifying and correcting misassembly errors. REAPR is designed to use both short insert and long insert paired-end sequencing libraries, however, it can operate with only one of these types of sequencing data. Amosvalidate, which is included as part of the AMOS assembly package [33], was developed specifically for first generation sequencing libraries [31]. iMetAMOS [34] is an automated assembly pipeline that provides error correction and validation of the assembly. It packages several open-source tools and provides annotated assemblies that result from an ensemble of tools and assemblers. Currently, it uses REAPR for misassembly error correction.

Many optical mapping tools exist and deserve mentioning, including AGORA [35], SOMA [36], and Twin [30]. AGORA [35] uses the optical map information to constrain de Bruijn graph construction with the aim of improving the resulting assembly. SOMA [36] uses dynamic programming to align *in silico* digested contigs to an optical map. Twin [30] is an index-based method for aligning contigs to an optical map. Due to its use of an index data structure it is capable of aligning *in silico* digested contigs orders of magnitude faster than competing methods. Xavier et al. [37] demonstrated misassembly errors in bacterial genomes can be detected using proprietary software.

Lastly, there are special purpose tools that have some relation to MISSEQUEL in their algorithmic approach. Numerous assembly tools use a finishing process after assembly, including Hapsembler [38], LOCAS [39], Meraculous [40], and the “assisted assembly” algorithm [41]. Hapsembler [38] is a haplotype-specific genome assembly toolkit that is designed for genomes that are highly-polymorphic. Both RACA [42], and SCARPA [43] perform paired-end alignment to the contigs as an initial step, and thus, are similar to our algorithm in that respect.

2 Methods

MISSEQUEL can be broken down into four main steps: recruitment of reads to contigs; construction of the red-black positional de Bruijn graph; misassembly error prediction; and misassembly verification using optical mapping data. We explain each of these steps in detail in the following subsections.

2.1 Recruitment of Reads and Threshold Calculation

MISSEQUEL first aligns reads to contigs in order to identify regions that contain abnormal read alignments. Collapsed or expanded repeats will present as the read coverage being greater or lower than the expected genome coverage in the region that has been misassembled. Similarly, inversion and rearrangement errors will present as the alignment of the mate-pairs being rearranged. Figures 1 and 2 illustrate these unconventional read alignments. More specifically, this step consists of aligning all the (paired-end) reads to all the contigs and then calculating three thresholds, Δ_L , Δ_U and Γ . The range $[\Delta_L, \Delta_U]$ defines the acceptable read depth, and Γ defines the maximum allowable number of reads whose mate-pair aligns in an unconventional manner (e.g. inverted orientation). In order to calculate these thresholds, we consider all alignments of each read as opposed to just the best alignment of each read since misassembly errors frequently occur within repetitive regions

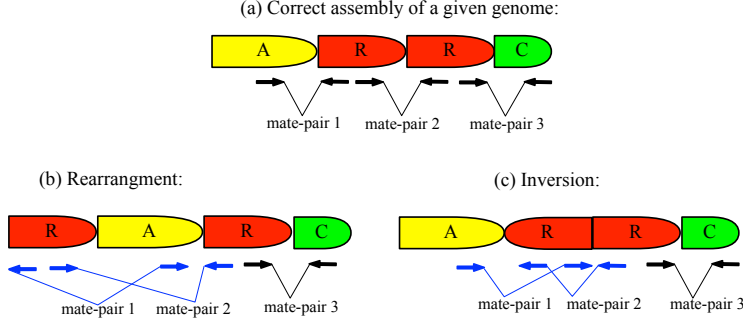


Fig. 1. An illustration of the systematic alterations that occur with rearrangements and inversions. (a) Shows the proper read alignment where mate-pair reads have the correct orientation and distance from each other. A rearrangement or inversion will present itself by the orientation of the reads being incorrect, and/or the distance of the mate-pairs being significantly smaller or larger than the expected insert size. This is shown in (b) and (c), respectively.

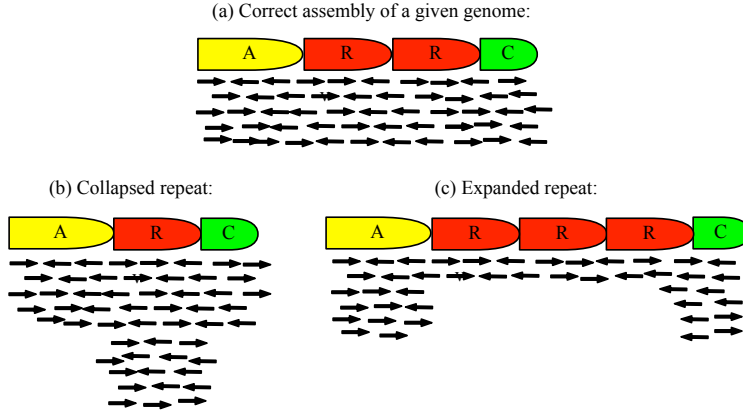


Fig. 2. An illustration of the systematic alterations that occur with collapsed or expanded repeats. (a) Shows the proper read depth, which is uniform across the genome. A collapsed or expanded repeat will manifest as significantly lower or higher read depth; (b) shows a collapsed repeat, where the read depth being significantly greater than expected; and (c) shows an expanded repeat, where the observed read depth is significantly lower than expected.

where the reads will align to multiple locations. MISSEQUEL performs this step using BWA (version 0.5.9) in paired-end mode with default parameters [44]. Subsequently, after alignment, each contig is treated as a series of consecutive 200 bp regions. These are sampled uniformly at random ℓ times, and the mean (μ_d) and the standard deviation (σ_d) of the read depth and the mean (μ_i) and the standard deviation (σ_i) of the number of alignments where an unconventional mate-pair orientation is witnessed are calculated from these sampled regions. Δ_L is set to the maximum of $\{0, \mu_d - 3\sigma_d\}$, Δ_U is set to $\mu_d + 3\sigma_d$, and Γ is set to $\mu_i + 3\sigma_i$. The default for ℓ is $\frac{1}{20}$ th of the contig length; this can be changed via an input parameter of MISSEQUEL.

2.2 Construction of the Red-Black Positional de Bruijn Graph

After threshold calculation, the red-black positional de Bruijn graph is constructed. For clarity, we begin by describing the *positional de Bruijn graph*, given by Ronen et al. [45], and then define the red-black positional de Bruijn graph. Whereas the edges in the traditional de Bruijn graph

correspond to k -mers, the edges in the positional de Bruijn graph correspond to k -mers and their inferred positions on the contigs (*positional k -mers*). Hence, the positional de Bruijn graph $G_{k,\Phi}$ is defined for a multiset of positional k -mers and parameter Φ , and is constructed in a similar manner to the traditional de Bruijn graph using an A-Bruijn graph framework from [46]. Given a k -mer s_k , let $\text{prefix}(s_k)$ be the first $k - 1$ nucleotides of s_k , and $\text{suffix}(s_k)$ be the last $k - 1$ nucleotides of s_k . Each positional k -mer (s_k, p) in the input multiset corresponds to a directed edge in the graph between two positional $(k - 1)$ -mers, $(\text{prefix}(s_k), p)$ and $(\text{suffix}(s_k), p + 1)$. After all edges are formed, the graph undergoes a gluing operation. A pair of positional $(k - 1)$ -mers, (s_{k-1}, p) and (s'_{k-1}, p') , are glued together into a single vertex if $s_{k-1} = s'_{k-1}$ and $p \in [p' - \Phi, p' + \Phi]$. Two positional $(k - 1)$ -mers are glued together if their sequences are the same and their positions are within Φ from each other. We refer to the *multiplicity* of a positional $(k - 1)$ -mer (s_{k-1}, p) as the number of occurrences where s_{k-1} clustered at position p .

MISSEQUEL constructs the red-black positional de Bruijn graph from the alignment of the reads to the contigs. The red-black positional de Bruijn graph contains positional k -mers and is constructed in an identical way as the positional de Bruijn graph with the addition that each vertex ($(k - 1)$ -mer) has an associated red or black color attributed to it that is defined using Δ_L , Δ_U and Γ . In addition to the multiplicity of each positional $(k - 1)$ -mer, the number of positional $(k - 1)$ -mers that originated from a read whose mate-pair did not align in the conventional direction is stored at each vertex. When the multiplicity is less than Δ_L or greater than Δ_U , or if the observed frequency of unconventional mate-pair orientation is greater than Γ , then the vertex is *red*; otherwise it is *black*.

2.3 Misassembly Conjecture and Breakpoint Estimation

A red-black positional de Bruijn graph is constructed for each contig, and misassembly errors in each contig are detected by searching for consecutive red vertices in the corresponding graph. Depth-first search is used for the graph traversal. If there are greater than 50 consecutive red vertices then the contig is conjectured to be misassembled. The breakpoint in the contig can be determined by recovering the position of the corresponding red vertices (e.g., the positional $(k - 1)$ -mers). The number of consecutive red vertices needed to consider it misassembled can be changed via a command line parameter in MISSEQUEL. Our experiments were performed with the default (e.g. 50), which corresponds to a region in the contig that has length ≥ 50 bp. After this stage of the algorithm, we take contigs having regions exceeding that threshold as a set of contigs that are conjectured to be misassembled and their transitions in and out of those regions as breakpoints.

2.4 Misassembly Verification

Lastly, we use optical mapping data to verify whether a contig that is conjectured to be misassembled indeed is. Verification is based on the expectation that, after *in silico* digestion, a correctly assembled contig has a sequence of fragment sizes that is similar to that in the optical map at the corresponding locus in the genome. In other words, an *in silico* digested contig should align to some region of the optical map since both are derived from the same region in the genome. Conversely, since misassembled contigs are not faithful reconstructions of any part of the genome, when *in silico* digested, their sequence of fragments will likewise not have a corresponding locus in the optical map to which it aligns.

Optical maps contain measurement error at each fragment size so some criteria is needed to decide whether variation in fragment size of an *in silico* digested contig and that of an optical map at a particular locus is due to variation in the size of the physical fragments or a consequence of optical measurement error. Due to this ambiguity, and the necessary tolerances to ensure correctly assembled contigs align to the locus in the optical map, misassembled contigs may also align to loci in the optical map, which by coincidence have a fragment sequence similar to the contig within the threshold margin of error. While there are various sophisticated approaches to determining statistical significance of an alignment, such as by Sakar et al. [47], we use a χ^2 model discussed by Nagarajan et al. [36] and take the cumulative density function ≤ 0.85 as evidence of alignment, which we found to work well empirically.

In addition, a misassembled contig only fails to align to the optical map if the enzyme recognition sequence, and thus the cleavage sites, exist in the contig in a manner that disrupts a good alignment (e.g. a misassembled contig with an inverted segment may still align if cleavage sites flank the inverted segment). This implies that (a) some enzymes produce optical maps that have greater performance in identifying misassembly errors; and (b) alignment to the optical map is not as strong evidence for correct assembly as non-alignment to the optical map is for misassembly. This leads to the conclusion that an ensemble of optical maps (each made with a different enzyme) has a greater chance at revealing misassembly errors than a single optical map. Since acquiring three optical maps for one genome is reasonably accessible for many sequencing projects, the process of *in silico* digestion and alignment is repeated for three enzymes and the consensus of the alignment is taken over all of them, i.e., if two out of three times the contig did not align then it is deemed not to align (by the consensus). A contig is deemed to be misassembled if it fails to align. The alignment is performed using Twin [30] (with default parameters) and then these results are filtered according to the χ^2 model mentioned above. For our experiments, optical maps were simulated by *in silico* digesting reference genomes, adding normally distributed noise with a 150 bp standard deviation, and discarding fragments smaller than 700 bp.

3 Results

3.1 Datasets

Our first dataset consists of approximately 6.9 million paired-end 101 bp reads from the prokaryote genome *Francisella tularensis*, generated by Illumina Genome Analyzer (GA) IIX platform. It was obtained from the NCBI Short Read Archive (accession number SRR063416). The reference genome was also downloaded from the NCBI website (Reference genome NC_006570.2). The *Francisella tularensis* genome is 1,892,775 bp in length. As a measure of quality assurance, we aligned the reads to the *Francisella tularensis* genome using BWA (version 0.5.9) [44] with default parameters. We call a read *mapped* if BWA outputs an alignment for it and *unmapped* otherwise. Analysis of the alignments revealed that 97% of the reads mapped to the reference genome, representing an average depth of approximately $367\times$.

Our second dataset consists of approximately 31.3 million paired-end 100 bp reads from the loblolly pine (*Pinus taeda* L.) genome [48]. We downloaded the reference genome from the pine genome website [49] and simulated reads from the largest five hundred scaffolds from the reference using ART [50] (art illumina). ART was ran with parameters that simulated 100 bp paired end reads with 200 bp insert size and 50x coverage. The substitution error rate was reduced to one 10th of the default profile. The data for this experiment is available on the MISSEQUEL website. We

simulated an optical map using the reference genome for *Francisella tularensis* and loblolly pine since there is no publicly available one for these genomes.

Table 1. The performance comparison between major assembly tools on the *Francisella tularensis* dataset (1,892,775 bp, 6,907,220 reads, 101 bp) using QUAST in default mode [14]. All statistics are based on contigs no shorter than 500 bp. N50 is defined as the length for which the collection of all contigs of that length or longer contains at least half of the sum of the lengths of all contigs, and for which the collection of all contigs of that length or shorter also contains at least half of the sum of the lengths of all contigs. # unaligned is the number of contigs that did not align to the reference genome, or only partially aligned (part). Total is sum of the length of all contigs. MA is the number of (extensively) misassembled contigs. local MA is the total number of contigs that had local misassemblies. MA (bp) is the total length of the MA contigs. GF is the genome fraction percentage, which is the fraction of genome bases that are covered by the assembly. -rr and +rr denotes before and after repeat resolution, respectively.

Assembler	# contigs (# unaligned)	N50	Largest (bp)	Total (bp)	MA	local MA	MA (bp)	GF (%)
Velvet	358 (3 + 35 part)	7,377	39,381	1,762,202	11	36	84,965	92.09
SOAPdenovo	307 (3 + 31 part)	8,767	39,989	2,018,158	10	35	96,258	92.05
ABYSS	96 (1 part)	27,975	88,275	1,875,628	64	32	1,330,684	95.87
SPAdes (-rr)	102 (2 + 11 part)	25,148	87,449	1,788,634	11	30	258,309	92.81
SPAdes (+rr)	100 (2 + 17 part)	26,876	87,891	1,797,197	23	31	497,356	93.75
IDBA	109 (1 + 10 part)	23,223	87,437	1,768,958	10	31	221,087	92.64

Table 2. The performance comparison between major assembly tools on Loblolly pine (*Pinus taeda* L.) genome dataset (62,647,324 bp, 31.3 million reads, 100 bp) using QUAST in default mode [14].

Assembler	# contigs (# unaligned)	N50	Largest (bp)	Total (bp)	MA	local MA	MA (bp)	GF (%)
Velvet	13,327 (0)	1,740	10,823	51,851,131	0	0	0	62.21
SOAPdenovo	16,126 (0 + 1 part)	7,950	63,004	57,205,817	0	0	0	90.01
ABYSS	4,586 (16 + 89 part)	37,089	201,382	63,349,408	127	715	1,391,565	98.17
SPAdes (-rr)	20,671 (4 + 10 part)	4,809	44,993	45,079,764	7	11	65,079	81.30
SPAdes (+rr)	8,607 (7 + 102 part)	16,957	108,442	59,730,939	299	57	3,734,609	94.57
IDBA	22,409 (3 + 31 part)	3,990	40,213	49,765,854	61	200	292,769	79.03

We assembled both sets of reads with a wide variety of state-of-the-art assemblers. The versions used were those that were publicly available before or on September 1, 2014: SPAdes (version 3.1) [51]; Velvet (version 1.2.10) [9]; SOAPdenovo (version 2.04) [10]; ABySS (version 1.5.2) [11]; and IDBA-UD (version 1.1.1) [52]. SPAdes outputs two assemblies: before repeat resolution and after repeat resolution — we report both. Some of the assemblers emitted both contigs and scaffolds. We considered contigs only but note that all scaffolds had a greater number of misassembly errors. *We emphasize that our purpose here is not to compare the various assemblers, but demonstrate that all assemblers produce misassembly errors, which are in need of consideration and correction.*

We used Quast [14] in default mode to evaluate the assemblies. Quast defines misassembly error as being *extensive* or *local*. A (extensive) misassembled contig is defined as one that satisfies one following conditions: (a) the left flanking sequence aligns over 1 kbp away from the right flanking sequence on the reference; (b) flanking sequences overlap on more than 1 kbp; (c) flanking sequences align to different strands or different chromosomes. Whereas, a local misassembled contig is one that satisfies the following conditions: (a) two or more distinct alignments cover the breakpoint;

(b) the gap between left and right flanking sequences is less than 1 kbp; and the left and right flanking sequences both are on the same strand of the same chromosome of the reference genome. We made a minor alteration to Quast to output which contigs contain local misassembly errors. A contig can contain both extensive and local misassembly errors. Any correctly assembled contig is one that does not contain either type of error.

3.2 Detection of Misassembly Errors in *Francisella tularensis*

Table 1 gives the assembly statistics corresponding to this experiment. Comparable assembly results on this data were reported by Ilie et al. [53], though in some cases we used more recent software releases (e.g., for SPAdes). Note that the number of locally misassembled contigs and the number of extensively misassembled contigs is not disjoint. A contig can be locally and extensively misassembled. Thus, Table 1 gives the number of contigs having at least one extensive misassembly error, and the number of contigs having at least one local misassembly error.

Table 3. The performance comparison of our method on the *Francisella tularensis* dataset. The true positive rate (TPR) in this context is a contig that is misassembled and is predicted to be so. The false positive rate (FPR) is a correctly assembled contig that was predicted to be misassembled. The TPR and FPR is given as a percentage with the raw values given in brackets

Correction Method	Assembler	MA TPR	local MA TPR	FPR
misSEQuel (paired-end data only)	Velvet	100% (11 / 11)	100% (36 / 36)	58% (180 / 312)
	SOAPdenovo	100% (10 / 10)	100% (35 / 35)	63% (165 / 263)
	ABYSS	100% (64 / 64)	100% (32 / 32)	87% (20 / 23)
	SPAdes (-rr)	100% (11 / 11)	100% (30 / 30)	83% (52 / 63)
	SPAdes (++rr)	100% (23 / 23)	100% (31 / 31)	86% (49 / 57)
	IDBA	100% (10 / 10)	100% (31 / 31)	38% (57 / 149)
misSEQuel (optical mapping data only)	Velvet	55% (6 / 11)	69% (25 / 36)	24% (76 / 312)
	SOAPdenovo	80% (8 / 10)	63% (22 / 35)	29% (77 / 263)
	ABYSS	69% (44 / 64)	88% (28 / 32)	13% (3 / 23)
	SPAdes (-rr)	91% (10 / 11)	87% (26 / 30)	21% (13 / 63)
	SPAdes (++rr)	87% (20 / 23)	81% (25 / 31)	16% (9 / 57)
	IDBA	90% (9 / 10)	77% (24 / 31)	10% (15 / 149)
misSEQuel (paired-end and optical mapping data)	Velvet	55% (6 / 11)	100% (26 / 36)	22% (68 / 312)
	SOAPdenovo	80% (8 / 10)	84% (21 / 35)	20% (53 / 263)
	ABYSS	69% (44 / 64)	88% (28 / 32)	13% (3 / 23)
	SPAdes (-rr)	91% (10 / 11)	87% (26 / 30)	19% (12 / 63)
	SPAdes (++rr)	97% (20 / 23)	81% (25 / 31)	16% (9 / 57)
	IDBA	90% (9 / 10)	77% (24 / 31)	9% (14 / 149)
REAPR	Velvet	55% (6 / 11)	11% (4 / 36)	< 1% (2 / 312)
	SOAPdenovo	20% (2 / 10)	14% (5 / 35)	2% (6 / 263)
	ABYSS	13% (8 / 64)	13% (4 / 32)	4% (1 / 23)
	SPAdes (-rr)	27% (3 / 11)	27% (8 / 30)	5% (3 / 63)
	SPAdes (++rr)	0% (0 / 23)	19% (6 / 31)	11% (6 / 57)
	IDBA	40% (4 / 10)	13% (4 / 31)	4% (6 / 149)

Table 3 shows the results for: (a) MISSEQUEL with paired-end data only; (b) MISSEQUEL with optical mapping data only; and (c) MISSEQUEL with both optical mapping and paired-end

data in order to demonstrate the gain of combining both types of data. As demonstrated by these results, using short read paired-end data alone produces a high false positive rate, since it is unable to distinguish between structural variations within the genome and misassembly errors. This is an inherent shortcoming of short read data and demonstrates that in order to decrease the false positive rate, another source of information must be used in combination. Optical mapping data has a much lower false positive rate and when used in combination with paired-end data, produces optimal results. The lowest false positive rate was witnessed when both optical mapping and paired-end data were used. In some cases, the reduction in the false positive rate was dramatic; from 87% (ABYSS, paired-end data) to 13% (ABYSS, paired-end and optical mapping data). The true positive rate of locally misassembled contigs was between 77% and 100% when both paired-end and optical mapping data were used. Lastly, true positive rate of extensively misassembled contigs was between 55% and 100% when both paired-end and optical mapping data were used.

In our experiments, we iterate through combinations of three enzymes from the REBASE enzyme database [54] and use the set of enzymes that performed best. Our results demonstrate that with a good enzyme choice over half of all extensively misassembled contigs, and over 75% of locally misassembled contigs can be identified with only a 9%-22% false discovery rate.

3.3 Detection of Misassembly Errors in Loblolly Pine

The results for the loblolly pine are shown in Table 4. Both Velvet and SOAPdenovo produced zero misassembled contigs on this dataset, so we do not include them in Table 4. MISSEQUEL correctly identifies between 31% and 100% of extensively misassembled contigs, and between 57% and 73% of locally misassembled contigs. The false positive rate was between 0.6% and 43%. Although, REAPR has a lower false positive rate (between 3% and 11%), it is only capable of identifying a small number of extensively misassembled contigs (between 2% and 14%) and a small number of locally misassembled contigs (between 2% and 27%).

Lastly, the restriction enzymes used in our experiments were chosen to be optimal by considering the set of all possible enzymes in the aforementioned database. Nonetheless, we note that if the enzyme combination was chosen at random then the expected false positive rate and true positive rate would decrease by a small fraction for majority of the assemblies considered. See the Appendix for prototypical ROC curves and heatmaps illustrating the density of enzyme combinations at various detection rates.

Table 4. The performance comparison of our method on the loblolly pine dataset. Again, a true positive in this context is a contig that is misassembled and is predicted to be so. A false positive is a correctly assembled contig that was predicted to be misassembled.

Correction Method	Assembler	MA TPR	local MA TPR	FPR
misSEQUEL	ABYSS	31% (40 / 127)	57% (405 / 715)	43% (1,604 / 3,754)
	SPAdes (-rr)	100% (7 / 7)	73% (8 / 11)	<1% (135 / 20,653)
	SPAdes (+rr)	67% (199 / 299)	67% (38 / 57)	38% (3,117 / 8,254)
	IDBA	52% (32 / 61)	73% (145 / 200)	19% (4,258 / 22,150)
REAPR	ABYSS	7% (9 / 127)	2% (12 / 715)	3% (112 / 3,754)
	SPAdes (-rr)	14% (1 / 7)	27% (3 / 11)	6% (1,323 / 20,653)
	SPAdes (+rr)	7% (21 / 299)	5% (3 / 57)	5% (424 / 8,254)
	IDBA	2% (1 / 61)	6% (12 / 200)	11% (2,354 / 22,150)

3.4 Practical Considerations: Memory and Time

We evaluated the memory and time requirements of MISSEQUEL. Since MISSEQUEL is a multi-threaded application, its wall-clock-time depends on the computing resources available to the user. MISSEQUEL required a maximum of 8 threads, 16 GB and 1.5 hours on all assemblies of *Francisella tularensis*, and a maximum of 20 GB and 2.5 hours to complete on all assemblies of loblolly pine. Most genome assemblers require an incomparably greater amount of time and memory and thus, from a practical perspective, the requirements of MISSEQUEL are not a significant increase. The difference in the resource requirements of MISSEQUEL in comparison to modern assemblers is due to the fact it operates contig-wise rather than genome-wise and therefore, only deals with a significantly smaller portion of the data at a single time. We conclude by mentioning that MISSEQUEL is not optimized for memory and time and both could be further reduced but reimplementing the red-black positional de Bruijn graph using memory and time succinct data structures.

4 Discussion and Conclusions

This paper describes the first non-proprietary computational method for identifying misassembly errors using short read sequence data and optical mapping data. Our results demonstrate: (1) a substantial number of misassembly errors can be identified in draft genomes of prokaryote and eukaryote species; (2) our method scales to large genomes; and (3) it can be used in combination with any assembler and thus, making it a viable post-processing step for any assembly.

While MISSEQUEL is capable of identifying a significant percentage of misassembly errors, it does not address the reassembly of those the misassembled contigs. Correcting misassembly errors by segmenting the contigs at their breakpoints will remove the errors but will also reduce the N50 of the assembly. For this reason, we believe that creating a reassembly tool to correctly reassemble contigs using the misassembly information and data warrants future investigation.

While our main contributions are the computational method itself and the demonstration that optical mapping can have significant benefit for misassembly detection, optimal results are contingent upon good enzyme selection. Thus, we conclude by suggesting that efficient algorithmic selection of enzymes that will yield such informative optical maps in a *de novo* scenario is an area for interesting and important future work.

Acknowledgements

The authors would like to thank Pavel Pevzner from the University of California, San Diego and Anton Korobeynikov from Saint Petersburg State University for many insightful discussions. In addition, we would like to thank Alexey Gurevich from Saint Petersburg State University for clarifications and support with Quast.

MM and CB were funded by the Colorado Clinical and Translational Sciences Institute which is funded by National Institutes of Health (NIH-NCATS, UL1TR001082, TL1TR001081, KL2TR001080). SJP was supported by the Helsinki Institute of Information Technology (HIIT) and by Academy of Finland through grants 258308 and 250345 (CoECGR). All authors greatly appreciate the funding provided for this project.

References

1. D. Haussler, S. J. O'Brien, O.A. Ryder, F.K. Barker, M. Clamp, A. J. Crawford, R. Hanner, O. Hanotte, W.E. Johnson, J.A. McGuire, et al. Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10,000 Vertebrate Species. *The Journal of Heredity*, 100(6):659–674, 2008.
2. G E Robinson, K J Hackett, M Purcell-Miramontes, S J Brown, J D Evans, M R Goldsmith, D Lawson, J Okamura, H M Robertson, and D J Schneider. Creating a Buzz About Insect Genomes. *Science*, 331(6023):1386–1386, 2011.
3. S. Ossowski, K. Schneeberger, R.M. Clark, C. Lanz, N. Warthmann, and D. Weigel. Sequencing of Natural Strains of *Arabidopsis thaliana* with Short Reads. *Genome Research*, 18:2024–2033, 2008.
4. P. J. Turnbaugh, R. E. Ley, M. Hamady, C. Fraser-Liggett, R. Knight, and J. I. Gordon. The Human Microbiome Project: Exploring the Microbial Part of Ourselves in a Changing World. *Nature*, 449(7164):804, 2007.
5. R.M Idury and M.S. Waterman. A New Algorithm for DNA Sequence Assembly. *Journal of Computational Biology*, 2(2):291–306, 1995.
6. P.A. Pevzner, H. Tang, and M.S. Waterman. An Eulerian Path Approach to DNA Fragment Assembly. *Proceedings of the National Academy of Sciences*, 98(17):9748–9753, 2001.
7. P.E.C. Compeau, P.A. Pevzner, and G. Tesler. How to Apply De Bruijn Graphs to Genome Assembly. *Nature Biotechnology*, 29(11):987–991, 2011.
8. M.J. Chaisson and P.A. Pevzner. Short Read Fragment Assembly of Bacterial Genomes. *Genome Research*, 18(2):324–330, 2008.
9. D.R. Zerbino and E. Birney. Velvet: Algorithms for *De Novo* Short Read Assembly using De Bruijn Graphs. *Genome Research*, 18(5):821–829, 2008.
10. R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, and J. Wang. *De Novo* Assembly of Human Genomes with Massively Parallel Short Read Sequencing. *Genome Research*, 20(2):265–272, 2010.
11. J.T. Simpson, K. Wong, S.D. Jackman, J.E. Schein, S.J. M. Jones, and I. Birol. ABySS: A Parallel Assembler for Short Read Sequence Data. *Genome Research*, 19(6):1117–1123, 2009.
12. J. Butler, I. MacCallum, M. Kleber, I.A. Shlyakhter, M.K. Belmonte, E.S. Lander, C. Nusbaum, and D.B. Jaffe. ALLPATHS: *De Novo* Assembly of Whole-Genome Shotgun Microreads. *Genome Research*, 18(5):810–820, 2008.
13. S.L. Salzberg. Beware of Mis-assembled Genomes. *Bioinformatics*, 21(24):4320–4321, 2005.
14. A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler. QUAST: Quality Assessment Tool for Genome Assemblies. *Bioinformatics*, 29(8):1072–1075, 2013.
15. D. C. Schwartz, X. Li, L. I. Hernandez, S. P. Ramnarain, E. J. Huff, and Y. K. Wang. Ordered Restriction Maps of *Saccharomyces Cerevisiae* Chromosomes Constructed by Optical Mapping. *Science*, 262:110–114, 1993.
16. C. Aston and D.C. Schwartz. *Optical Mapping in Genomic Analysis*. John Wiley and Sons, Ltd, 2006.
17. Dimalanta et al. A Microfluidic System for Large DNA Molecule Arrays. *Analytical Chemistry*, 76(18):5293–5301, 2004.
18. R. K. Neely, J. Deen, and J. Hofkens. Optical Mapping of DNA: Single-Molecule-Based Methods for Mapping Genome. *Biopolymers*, 95(5):298–311, 2011.
19. T. Anantharaman and B. Mishra. A probabilistic Analysis of False Positives in Optical Map Alignment and Validation. In *Proceedings of WABI*, pages 27–40, 2001.
20. B. Teague et al. High-Resolution Human Genome Structure by Single-Molecule Analysis. *Proceedings of the National Academy of Sciences*, 107(24):10848–10853, 2010.
21. S. Reslewic et al. Whole-Genome Shotgun Optical Mapping of *Rhodospirillum Rubrum*. *Applied and Environmental Microbiology*, 71(9):5511–5522, 2005.
22. S. Zhou et al. A Whole-Genome Shotgun Optical Map of *Yersinia pestis* Strain KIM. *Applied and Environmental Microbiology*, 68(12):6321–6331, 2002.
23. S. Zhou et al. Shotgun Optical Mapping of the Entire *Leishmania major* Friedlin Genome. *Molecular and Biochemical Parasitology*, 138(1):97–106, 2004.
24. S. Zhou et al. Validation of Rice Genome Sequence by Optical Mapping. *BMC Genomics*, 8(1):278, 2007.
25. S. Zhou et al. A Single Molecule Scaffold for the Maize Genome. *PLoS Genetics*, 5(11):e1000711, 11 2009.
26. D. M. Church et al. Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse. *PLoS Biology*, 7(5):e1000112+, 2009.
27. Y. Dong et al. Sequencing and Automated Whole-Genome Optical Mapping of the Genome of a Domestic Goat. *Nature Biotechnology*, 31(2):136–141, 2013.
28. J. T. Howard et al. *De Novo* High-coverage Sequencing and Annotated Assemblies of the *Budgerigar* Genome, 2014.

29. S. Chamala et al. Assembly and Validation of the Genome of the Nonmodel Basal Angiosperm *Amborella*. *Science*, 342(6165):1516–1517, 2013.
30. M. Muggli, S. Puglisi, and C. Boucher. Efficient Indexed Alignment of Contigs to Optical Maps. In *Proceedings of the 14th Workshop on Algorithms in Bioinformatics (WABI 2014)*, pages 68–81, 2014.
31. A. Phillippy, M. Schatz, and M. Pop. Genome Assembly Forensics: Finding the Elusive Mis-Assembly. *Genome Biology*, 9(3):R55+, 2008.
32. M. Hunt, T. Kikuchi, M. Sanders, C. Newbold, M. Berriman, and T. D. Otto. REAPR: A Universal Tool for Genome Assembly Evaluation. *Genome Biology*, 14:R47, 2013.
33. Todd J. Treangen, Dan D. Sommer, Florent E. Angly, Sergey Koren, and Mihai Pop. *Next Generation Sequence Assembly with AMOS*, volume Chapter 11. John Wiley & Sons, Inc., 2011.
34. Sergey Koren, Todd J. Treangen, Christopher M. Hill, Mihai Pop, and Adam M. Phillippy. Automated ensemble assembly and validation of microbial genomes. *BMC Bioinformatics*, 15(126), 2014.
35. H.C Lin, S. Goldstein, L. Mendelowitz, S. Zhou, J. Wetzel, D.C. Schwartz, and M. Pop. AGORA: Assembly Guided by Optical Restriction Alignment. *BMC Bioinformatics*, 12:189, 2012.
36. N. Nagarajan, T.D. Read, and M. Pop. Scaffolding and Validation of Bacterial Genome Assemblies using Optical Restriction Maps. *Bioinformatics*, 24(10):1229–1235, 2008.
37. B.B. Xavier, J. Sabirova, M. Pieter, J.-P. Hernalsteens, H. de Greve, H. Goossens, and S. Malhotra-Kumar. Employing whole genome mapping for optimal *de novo* assembly of bacterial genomes. *BMC Research Notes*, 7:484, 2014.
38. N. Donmez and M. Brudno. Hapsembler: An Assembler for Highly Polymorphic Genomes. In *Proceedings of the 15th Annual international conference on Research in Computational Molecular Biology (RECOMB)*, pages 38–52, 2011.
39. J.D. Klein, S. Ossowski, K. Schneeberger, D. Weigel, and D.H. Huson. LOCAS—A Low Coverage Assembly Tool for Resequencing Projects. *PloS ONE*, 6(8):e23455+, 2011.
40. J.A. Chapman, I. Ho, S. Sunkara, S. Luo, G.P. Schroth, and D.S. Rokhsar. Meraculous: *De Novo* Genome Assembly with Short Paired-End Reads. *PLoS ONE*, 6(8):e23501, 2011.
41. S. Gnerre, E. Lander, K. Lindblad-Toh, and D. Jaffe. Assisted Assembly: How to Improve a *De Novo* Genome Assembly by Using Related Species. *Genome Biology*, 10(8):R88, 2009.
42. J. Kim et al. Reference-Assisted Chromosome Assembly. *Proceedings of the National Academy of Sciences*, 110(5):1785–1790, 2013.
43. N. Donmez and M. Brudno. SCARPA: Scaffolding Reads with Practical Algorithms. *Bioinformatics*, 29(4):428–434, 2013.
44. H. Li and R. Durbin. Fast and Accurate Short Read Alignment with Burrows–Wheeler Transform. *Bioinformatics*, 25(14):1754–1760, 2009.
45. R. Ronen, C. Boucher, H. Chitsaz, and P. Pevzner. SEQuel: Improving the Accuracy of Genome Assemblies. *Bioinformatics*, 28(12):i188–i196, 2012.
46. P.A. Pevzner, H. Tang, and G. Tesler. *De Novo* Repeat Classification and Fragment Assembly. *Genome Research*, 14(9):1786–1796, 2004.
47. Deepayan Sarkar, Steve Goldstein, David C Schwartz, and Michael A Newton. Statistical significance of optical map alignments. *Journal of Computational Biology*, 19(5):478–492, 2012.
48. D. Neale et al. Decoding the Massive Genome of Loblolly Pine using Haploid DNA and Novel Assembly Strategies. *Genome Biology*, 15(3):R59+, 2014.
49. PINEREFSEQ. <http://pinegenome.org/pinerefseq>. Accessed: 2014-09-01.
50. W. Huang, L. Li, J.R. Myers, and G.T. Marth. ART: A Next-Generation Sequencing Read Simulator. *Bioinformatics*, 28(4):593–594, 2012.
51. A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, A.S. Kulikov, V.M. Lesin, S.I. Nikolenko, S. Pham, A. D. Prjibelski, A.V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M.A. Alekseyev, and P. A. Pevzner. SPAdes: A New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, 19(5):455–477, 2012.
52. Y. Peng, H. C. Leung, S. M. Yiu, and F. Y. Chin. IDBA-UD: A *De Novo* Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth. *Bioinformatics*, 28(11):1420–1428, 2012.
53. L. Ilie, B. Haider, M. Molnar, and R. Solis-Oba. SAGE: String-overlap Assembly of GENomes. *BMC Bioinformatics*, 15:302, 2014.
54. J Roberts Richard, Vincze Tamas, Posfai Janos, and Macelis Dana. Rebase—a database for dna restriction and modification: Enzymes, genes and genomes. *Nucleic Acids Res*, 38:D234–D236, 2010.

5 Appendix

Figures 3 and 4 are the ROC plots for all (139 choose 3) enzymes for SOAPdenovo and IDBA assembly of *Francisella tularensis*, respectively. The heat maps give an idea of the probability of getting a particular true positive rate and false positive rate with a specific choice of enzymes. These plots show the sensitivity and specificity of misassembly detection using optical mapping data alone. The paired-end sequence data was not used. As can be seen in the plots, if a set of enzymes were chosen at random then optical mapping would still be informative and produce a meaningful classifier.

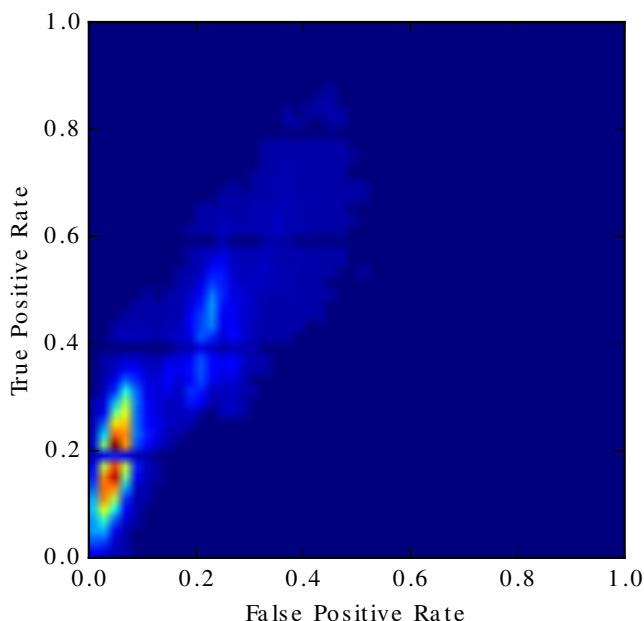


Fig. 3. ROC plot illustrating the density of optical map alignment based missassembly detection classification rates for the SOAPdenovo assembly of *Francisella tularensis*. The color intensity at each point indicates the number of three enzyme based classifiers having that classification rate. The plot includes results for optical maps with all three enzyme combinations using a set of 135 enzymes randomly drawn from the REBASE database. The velvet assembly (which is not shown) has a similar pattern. Hot spots represent the likely classification rate for enzymes chosen at random.

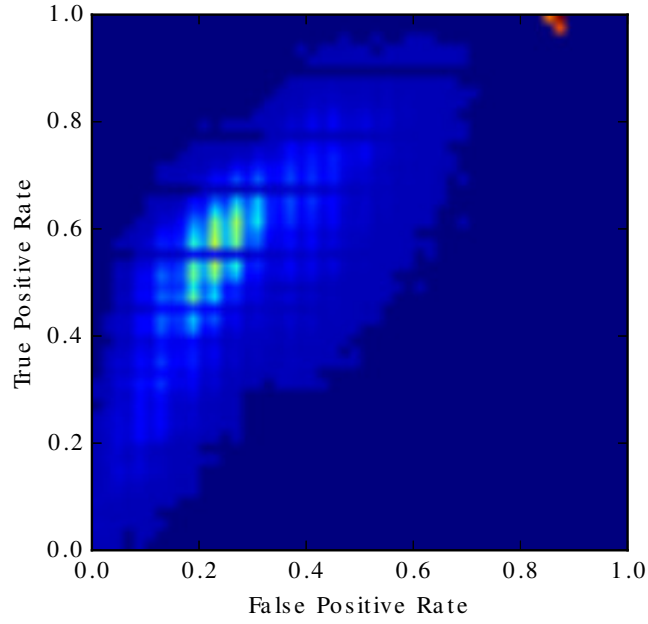


Fig. 4. ROC plot illustrating the density of optical map alignment based missassembly detection classification rates for the IDBA assembly of *Francisella tularensis*. The color intensity at each point indicates the number of three enzyme based classifiers having that classification rate. The plot includes results for optical maps with all three enzyme combinations using a set of 135 enzymes randomly drawn from the REBASE database. Both SPAdes assemblies as well as ABySS (which are not shown) have a similar pattern. Hot spots represent the likely classification rate for enzymes chosen at random.